The visual system supports on-line translation invariance for object identification.

Jeffrey S. Bowers

Ivan I. Vankov

Casimir J.H. Ludwig

School of Experimental Psychology

University of Bristol,

12a Priory Road

Bristol, BS8-1TU

Key terms: translation invariance; translation tolerance; object identification; vision

Abstract:

The ability to recognize the same image projected to different retinal locations is critical for visual object recognition in naturalist contexts. On many theories translation invariance for objects only extends to trained retinal locations. On this approach, a familiar object projected to a non-trained location should not be identified. On another approach invariance is achieved "on-line", such that learning to identify an object in one location immediately affords generalization to other locations. We trained participants to name novel objects at one retinal location using eye-tracking technology and then tested their ability to name the same images presented at novel retinal locations. Across three experiments we found robust generalization. These findings provide a strong constraint for theories of vision.

Introduction Retinal images vary when an object is seen under different viewing conditions, including changes in orientation, viewing distance, illumination, and position in the visual field. Somehow the visual system must ensure that object recognition is invariant to these changes. Here we focus on how the visual system copes with translation across retinal position. Can we recognize an object that we have only ever experienced in one part of the visual field in a novel retinal location? For object recognition to be tolerant to changes in retinal position, do we need to have experienced the object in all possible retinal locations?

There are three general views regarding how we achieve translation tolerance (see Figure 1). On hypothesis 1, tolerance is largely post-visual. That is, our ability to identify familiar objects across a wide range of eccentricities is achieved by learning multiple high-level object representations for the same object at different retinal locations and then linking these representations to a common post-visual code, as in Figure 1A (e.g., Afraz & Cavanagh, 2008; Dehaene, Cohen, Sigman, & Vinckier, 2005; Kravitz, Kriegeskorte, & Baker, 2010; Ullman, 2007). On hypothesis 2, robust translation tolerance occurs within the visual system. That is, there is a common high-level visual representation for a given object, but contacting this representation requires training of the mapping between an object in different retinal locations and its higher-level visual representation, as depicted in Figure 1B (e.g., Cox & DiCarlo, 2008; Dandurand, Hannagan, & Grainger, 2013; Di Bono & Zorzi, 2013). A critical prediction of both hypotheses is that an object cannot be identified when it is projected to a novel retinal location that is distal from "trained" locations. Finally, on hypothesis 3, translation tolerance occurs within the visual system and is computed on-line. That is, the visual system maps a given object projected to different retinal locations to a common (single) high-level visual representation regardless of its retinal location. An object can be identified at a novel location, even when this new location is quite distal from the locations in which the object was trained (e.g., Biederman, 1987).

The claim that translation tolerance is computed on-line is often described the standard view (e.g., Kravitz, Kriegeskorte, & Baker, 2010). Early research appeared to support this view. For example, Biederman and Cooper (1991) showed that long-term priming for familiar objects is equally robust following a study-to-test change in retinal location. Similarly, in masked priming studies with words, robust and equal priming has been observed when the prime and targets are presented at the same or different retinal locations (Bowers & Turner, 2005). An alternative account of these findings, however, is that priming reflected post-visual processes, such as common name codes (Kravitz et al., 2010). Indeed, a number of studies designed to minimize the role of post-visual processing failed to observe robust (or any) translation tolerance in priming and perceptual learning tasks (e.g., Dill & Fahle, 1997; McAuliffe & Knowlton, 2000; Newell, Sheppard, Edelman, & Shapiro, 2005; Kravitz et al., 2010; for review see Kravitz, Vinson, & Baker, 2008). A common conclusion from these and related studies is to reject the on-line account of tolerance and instead adopt hypotheses 1 or 2.

However, there are a number of design features of these later studies that make it difficult to draw any strong conclusions. First, the objects are typically briefly flashed at a given retinal location at study and/or test (e.g., Dill & Fahle, 1997; Kravitz et al., 2010; McAuliffe & Knowlton, 2000), and it is possible that tolerance requires more extended sampling of the object. Indeed, other invariances, including left-right reflection and picture-plane rotations invariance, are only manifest in priming tasks when participants were able to attend to the objects for longer

periods of time (e.g, Thoma, Davidoff, & Hummel, 2007). Second, many of these studies have used images that are highly unlike real objects (e.g., Dill & Fahle, 1997), or images that are extremely similar (e.g., Cox & DiCarlo, 2008). Distinguishing between highly similar patterns may rely more heavily on low-level visual representations (Ahissar & Hochstein, 2004), and low-level codes are more retinotopically constrained.

In order to provide a strong test of on-line translational tolerance it is necessary to adopt conditions in which (i) more extended sampling is possible; (ii) the items are more object-like; (iii) the objects differ from one another in more than some fine perceptual detail; and (iv) post-visual codes cannot contribute to performance. To this end we used eye-tracking so that the objects could be presented for a longer duration at controlled retinal locations. In addition, participants were trained on a set of novel objects that differed in configurable properties rather than fine details. The question is whether participants could identify these objects in novel retinal locations after training in one location.

We used a naming task that requires the unique identification of the objects. The naming task does engage post-visual processes, but these codes cannot support translation tolerance in our experiments given that the objects were novel. To see why, consider a newly learned visual object representation that is tightly bound to a specific retinal location (as in hypothesis 1 and 2). By definition, this representation cannot be accessed when the object is projected to very different retinal locations. And if the newly learned visual representation cannot be accessed then the objects cannot be named given that the name codes are linked to these object codes (see Figure 1). This contrasts with previous Biederman studies (e.g., Biederman & Cooper, 1991; Fiser & Biederman, 2001) that assessed translation tolerance for familiar objects. In this situation, post-visual codes are contacted regardless of the location of the object at study and test (e.g., the semantics and the name of piano can be accessed wherever the piano is projected on retina given the participants' past history with pianos), and accordingly, post-visual cues could indeed have contributed to priming in these studies even if visual representations are tightly bound to retinal position.

Experiment 1

Materials and methods

Participants Ten participants took part in Experiment 1a and another ten participated in Experiment 1b. All participants were paid £10 for their time.

Stimuli and equipment We took six objects from Tarr and Pinker (1990) that include similar local features but that differ in their overall configuration (see Figure 2). We chose these objects because they cannot be identified or distinguished from one another on the basis of their parts but need to be identified as complete objects. No object was the mirror image of another, and they were all non-symmetrical along the vertical plane. Each object was assigned a spoken name ('Q', 'V', 'C', 'S, 'D', 'J').

Presentation of the stimuli was managed by a MATLAB (MathWorks, Inc.) program using Psychtoolbox 3.08. The objects were displayed on ViewSonic G225f 21-inch CRT monitor, running at 85 Hz with a spatial resolution of 1,024 × 768 and viewed at a distance of 57cm with objects extending 5º x 5º of visual angle. In Experiment 1a, the horizontal eccentricity from the center of the object to a central fixation cross was 5.5º. Experiment 1b is a replication, with an increased separation between the study and test locations (6.5º) and also a tighter region around fixation (see below). Participants were instructed to focus on a central fixation cross while eye

movements were tracked at 1000Hz using the EyeLink 2k system (SR Research Ltd.). Each time the system detected that the participants' gaze moved sufficiently far away from fixation (more than 2º in Experiment 1a and 1.5º in Experiment 1b) it replaced the object with a mask that prompted the participant to return to the fixation point. The time during which a mask was shown was not included in the total presentation time of the stimulus.

Procedure and design In Experiment 1a and 1b the participants focused on a centrally located fixation cross while an object was displayed in either the left or the right part of the screen. The experiment started with 30 familiarization trials during which a random sequence of the novel objects shown in Figure 2 was presented along with their spoken names ('Q', 'V', 'C', 'S, 'D', 'J'), with the object-name mappings counterbalanced across participants. Participants were instructed to learn the object names. Then participants completed a training phase. On each training trial an object was displayed for 2 seconds and participants attempted to retrieve its name. Written feedback was then provided along with a repetition of the object and the object name. An example of training trial is shown in Figure 3. The training lasted until the participant managed to correctly name 24 novel objects in a row. The average number of trails needed to complete the training sessions was 112 in Experiment 1a and 92 in Experiment 1b. There was 1 person who did not manage to complete the training for 150 trials and was replaced by another participant. During familiarization and training, all the objects were presented either to the left or the right side of fixation, with location counterbalanced across participants.

Participants who successfully completed the training phase completed the test phase that included 54 test trials in which the position of the objects was manipulated. For 18 test trials the object was presented at the same position as in training ('same' condition), for 18 trials the objects was presented in the center of the screen ('center') and for 18 trials the object was presented at the opposite side of the screen ('opposite'). There was no feedback in the testing phase, and the order of the test trials was randomized.

Results

As can be seen in Figure 4, the participants were able to reliably name the trained objects when presented at novel positions, either in the center or at the opposite of the screen. For example, in Experiment 1a, the average accuracy rate was 92% in the *same* condition, 81% in *center*, and 71% in the *opposite* condition. These results are inconsistent with the critical predictions of both Hypothesis 1 and 2 according to which performance should be at chance (16.67%).

In all of the conditions, the average accuracy was far above the chance level (Cohen d's - same: 8.26, center 5.06, opposite: 3.69). The effect sizes of the differences between experimental conditions were lower: same-center: Cohen d=1.06, same-opposite: d=1.78, center-opposite: d=.73. A similar pattern of results was found in Experiment 1b, with performance far above chance in all conditions: same: Cohen d=6.54, center: d=4.28, opposite: d=2.27. Given the increased eccentricity of the objects and the stricter fixation conditions, this serves to highlight the robustness of the translation effects. The effect sizes of the differences between conditions were lower than in Experiment 1a: same-center: Cohen d=.46, same-opposite: d=1.01, center-opposite: d=.62.

In these two experiments both the retinal and the spatial location (in screen coordinates) of the objects differed between training and test. In Experiment 2 we assessed to what extent reduced performance was due to changes in the spatial location. Objects were always presented in the centre of the screen, but the location of

the fixation point varied between study and test. In this way, the retinal location varied, but the spatial location was held constant.

Experiment 2

Materials and methods

Participants Ten participants took part in each of the experiments and were paid £10 for their time.

3.1.2. Stimuli and equipment The same equipment and stimuli were used as above. The horizontal eccentricity of the fixation point was 6.5º, and each time that the participants' gaze was more than 1.5º away from fixation it replaced the object with a mask that prompted the participant to return to the fixation point.

Procedure and design. The experimental procedure was similar to above. The only difference was that, during training, the novel objects were centered at the middle of the screen and the fixation point was located either to the left or right side of the screen, with location counterbalanced across participants. During testing, the fixation point was displayed either at the training position (*same*), at the center of the screen (*center*) or at the opposite side (*opposite*) in a randomized order. The average number of trials to complete the training session was 114.

Results. The results of Experiment 2 are also shown in Figure 4. The pattern is similar to what was found in Experiments 1ab. The accuracy was above chance in all the experimental conditions: *same*: Cohen d=4.86; *center:* 89%, d=4.29; *opposite*: 77%, d=4.45. Performance was better in the same compared to the opposite condition, d=.96, but not compared to the center condition, d=.07. There was also a significant reduction between the center and opposite conditions, d=.83.

In the final experiment we attempted to eliminate the effect of retinal location by presenting four of the six objects in multiple retinal locations during study with the remaining two objects projected to only one location. Previous work on perceptual learning suggests that tolerance may be improved when a number of similarly complex objects are experienced at multiple retinal locations (e.g. Xiao et al., 2008).

Experiment 3

Materials and methods

Participants Ten participants took part and were paid £10 for their time.

Stimuli and equipment The same equipment and stimuli were used as above, but in this case, the three novel objects were side-by-side, with one to the left, one at the middle, and one to the right side of the screen. The fixation cross was presented in the middle position. The horizontal eccentricity of the objects presented in the periphery was 5.5º. Each time that the participants' gaze was more than 2º away from fixation it replaced the objects with a mask that prompted the participant to return to the fixation point.

Procedure and design

The experiment started with a familiarization phase in which triplets of novel objects were displayed and their names were presented auditorily, one by one, from left to right. The two critical objects were presented at one retinal location - one to the left of fixation, one to the right. The critical objects varied across participants, but the same names were used in all cases. The remaining four objects were presented at all positions. The familiarization phase was followed by a training phase in which the participants saw triplets of objects and named them left to right. Feedback was provided after each trial. A response was regarded as incorrect if any of three objects was named incorrectly. Training was completed when participants managed to respond correctly ten times in a row.

The experiment ended with a test phase in which all the objects were presented at all positions. Thus, four experimental conditions were formed: *same* (the object was trained in periphery only and it is tested at the trained position), *center* (the object was trained in the periphery and is tested in the center position), *opposite* (the object was trained in the periphery and is tested in the opposite peripheral position), *control* (the objects was trained at all positions). There was no feedback at test.

Results

The results of Experiment 3 are displayed in Figure 5. In all conditions, accuracy was far above the chance level: *same*: 95%, d=11.43; *center:* 80%, d=3.54; *opposite*: 80%, d = 2.85; *control*: 92%, d=9.45. However performance was reduced in the center compared to same condition, d=1.08, as well as in the opposite condition, d=.68. There was no difference in the *center* compared to *opposite* condition, d=.00.

Once again performance was reduced when objects were presented at different retinal locations during training and test. This highlights that the retinotopic contribution of object learning that is difficult to eliminate, and is observed even when the control novel objects were trained at multiple locations.

General Discussion

The results are clear-cut: After participants learned to name novel objects at one retinal location they were able to identify and name the same objects at other retinal locations with a high degree of accuracy. These results rule out all theories that assume that there is little translation tolerance for objects within the visual system (hypothesis 1) or that robust tolerance within the visual system requires training a given object at a given retinal location (hypothesis 2), as outlined in Figure 1. Instead, the results lend some support "on-line" theories of tolerance in which high-level objects codes are represented independently of retinal location such that generalization is possible following an encounter with an object in a single location (hypothesis 3).

What should be made of the drop in performances following a study-to-test retinal change? Does this lend some support to theories according to which visual objects codes are tightly bound to retinal location (hypothesis 1) or to theories that claim that tolerance needs to be explicitly trained (hypothesis 2)? Not at all. In fact, our findings are in striking contrast with past work that has been used to support these hypotheses. For instance, Cox and DiCarlo (2008) argued for hypothesis 2 based on the finding that a rhesus monkey was at chance at identifying a novel object following a translation from +2º to -2º from fixation (and the fact that the responses of IT neurons showed the same selectivity in two monkeys). By contrast, in our Experiment 3, participants were ~80% accurate in naming novel objects following a shift of 13º (when chance was 16.7%).

Furthermore, all theories of visual object identification agree that vision is mediated by a hierarchical system in which early representations are tightly bound by retinal location (e.g., simple cells in V1). Thus, the effect of location may reflect the fact that performance was supported, in part, by low-level visual learning that boosted performance in the same location condition. For instance, the feature ⌐ occurs in two of the novel objects (see Figure 2), and if participants learned to map this lower-level feature to the object names this could contribute to performance in a retinotopically-constrained fashion. Note, the low-level features of our objects did not reliably predict the name of the objects – it is the overall configuration of the feature that defined the objects – so it is not possible to explain the high accuracy in naming across conditions on this basis. Nevertheless, learning these low-level features might boost performance in the same condition, as we observed.

Alternatively, the reduced performance across locations might reflect a limitation of translation tolerance at the object level (with no contribution from low-level features), contrary to Hypothesis 3. On this hypothesis, translation tolerance is indeed limited, but the tolerance is much greater than commonly claimed (as in Hypothesis 1 and 2).

An obvious question remains; namely, why did we obtain robust translation tolerance when most previous work has reported much more restricted tolerance? We can only speculate, but there are many methodological differences between studies that may explain the contrasting results. For example, Cox and DiCarlo (2008) trained two monkeys to discriminate briefly flashed objects that differed in a subtle visual detail over the course of 30-60 training sessions in which each object was presented in each location over 20,000 times. The fact that objects were briefly presented at study and test, the fact that the objects only different in visual detail, and the fact that training was extended over such a long time may all have contributed to the results. Indeed, DiCarlo and Maunsell (2003) suggested that the extensive training of objects at specific retinal locations may narrow receptive fields in macaque IT.

Similarly, as noted earlier, many of the studies that reported little translation invariance out with humans also flashed objects at study and test and required participants to distinguish between objects that differed in visual detail. Again, these factors may have contributed to limited translation invariance. Indeed, recent studies with humans have highlighted how training conditions can dramatically impact on the translation tolerance of low-level perceptual learning (e.g., learning to discriminate between subtle variations in contrast and orientation), with some studies showing little tolerance (e.g., Karni & Sagi, 1991) and other showing robust tolerance (e.g., Xiao et al., 2008). Accordingly, we think it is likely that the different translation results obtained with objects reflect something about the study and test conditions rather than the subject population (e.g., monkey or human). The most critical point, however, is that the visual system can support robust translation tolerance under some conditions that are arguably more ecologically valid.

The current findings are important because they provide a challenge for most theories and models. For example, many neural network models of word and object identification that claim to support translation tolerance implement specific versions of hypothesis 1 or 2 (Dandurand et al., 2013; Di Bono & Zorzi, 2013). That is, tolerance was achieved by training the models with objects (in these cases words) in all possible spatial locations. What the authors did not test is whether their models could generalize and identify an object at an untrained location. Almost certainly the answer is “no” as no mechanisms were included in order to achieve this capacity. Our findings highlight the need to develop processes that can support more robust translation tolerance (e.g., Foldiak et al., 1991) that are omitted in most current theories and models of object and word recognition.

Acknowledgments:

This research was supported by Leverhulme Grant RJ5538, awarded to Jeffrey S. Bowers

References:

Afraz, S.-R., & Cavanagh, P. (2008). Retinotopy of the face aftereffect. *Vision Research*, *48*(1), 42–54.

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, *8*(10), 457–464.

Biederman, I., & Cooper, E. E. (1991). Evidence for complete translational and reflectional. *Perception*, *20*, 585-593.

Bowers, J. S., & Turner, E. L. (2005). Masked priming is abstract in the left and right visual fields. *Brain and Language*, *95*(3), 414–422.

Cox, D. D., & DiCarlo, J. J. (2008). Does Learned Shape Selectivity in Inferior Temporal Cortex Automatically Generalize Across Retinal Position? *Journal of Neuroscience*, *28*(40), 10045–10055.

Dandurand, F., Hannagan, T., & Grainger, J. (2013). Computational models of location-invariant orthographic processing. *Connection Science*, *25*(1), 1–26.

Davis, C. J. (2010). The Spatial Coding Model of Visual Word Identification. *Psychological Review*, *117*(3), 713–758. doi:10.1037/a0019738

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: a proposal. *Trends in Cognitive Sciences*, *9*(7), 335–341.

Di Bono, M.G., & Zorzi, M. (2013). Deep generative learning of location-invariant visual word recognition. *Frontiers in Psychology*, *4*. 635.

DiCarlo, J. J. & Maunsell, J.H. (2003). Anterior Inferotemporal Neurons of Monkeys Engaged in Object Recognition Can be Highly Sensitive to Object Retinal Position. *Journal of Neurophysiology*, *89*(6), 3264–3278.

Dill, M., & Fahle, M. (1997). The role of visual field position in pattern-discrimination learning. *Proceedings of the Royal Society B: Biological Sciences*, *264*(1384), 1031–1036.

Fiser, J., & Biederman, I. (2001). Invariance of long-term visual priming to scale, reflection, translation, and hemisphere. *Vision Research*, *41*(2), 221-234.

Karni, A., and Sagi, D. (1991). Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. Proceedings of the National Academy of Sciences, 88, 4966–4970.

Kravitz, D. J., Kriegeskorte, N., & Baker, C. I. (2010). High-Level Visual Object Representations Are Constrained by Position. *Cerebral Cortex*, *20*(12), 2916–2925.

Kravitz, D. J., Vinson, L. D., & Baker, C. I. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences*, *12*(3), 114–122.

McAuliffe, S. P., & Knowlton, B. J. (2000). Long-term retinotopic priming in object identification. *Perception & Psychophysics*, *62*(5), 953–959.

Newell, F. N., Sheppard, D. M., Edelman, S., & Shapiro, K. L. (2005). The interaction of shape- and location-based priming in object categorisation: Evidence for a hybrid "what plus where" representation stage. *Vision Research*, *45*(16), 2065–2080.

Tarr, M. J., & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science*, *1*(4), 253–256.

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences*, *11*(2), 58–64.

Xiao, L.-Q., Zhang, J.-Y., Wang, R., Klein, S. A., Levi, D. M., & Yu, C. (2008). Complete Transfer of Perceptual Learning across Retinal Locations Enabled by Double Training. *Current Biology*, *18*(24), 1922–1926.

Figure 1

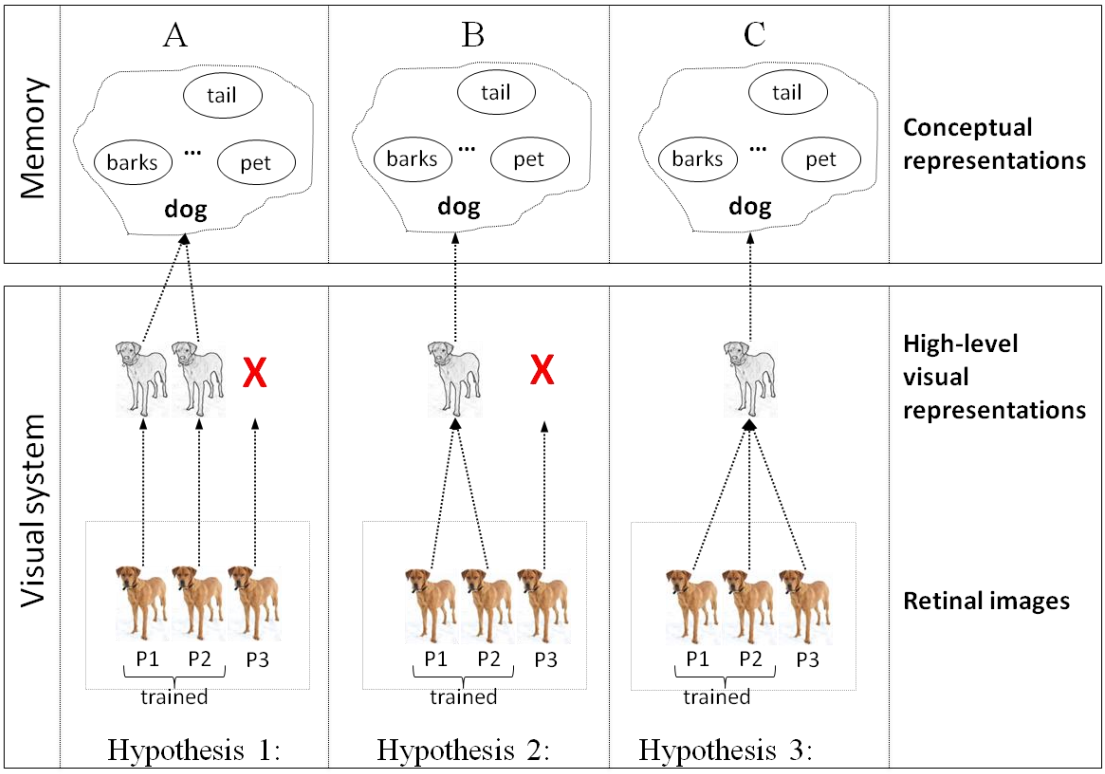

Illustration of three hypotheses regarding translation invariance. In Figure 1a translation invariance is largely post-visual. On this view, the visual system learns many high-level object representations for a given image, with different representations at different retinal locations. Translation invariance is achieved by mapping these distinct visual representations to common conceptual code. In Figure 1b translation invariance occurs within the visual system, but only at trained retinal locations. That is, the visual system learns to map a given image projected to different retinal locations onto a common (single) high-level visual perceptual code, but these mappings require training. On hypothesis 3, translation invariance occurs within the visual system and is computed "on-line". That is, the visual system maps a given image projected to different retinal locations to a common (single) high-level perceptual code, even when the image is projected to a novel retinal location.

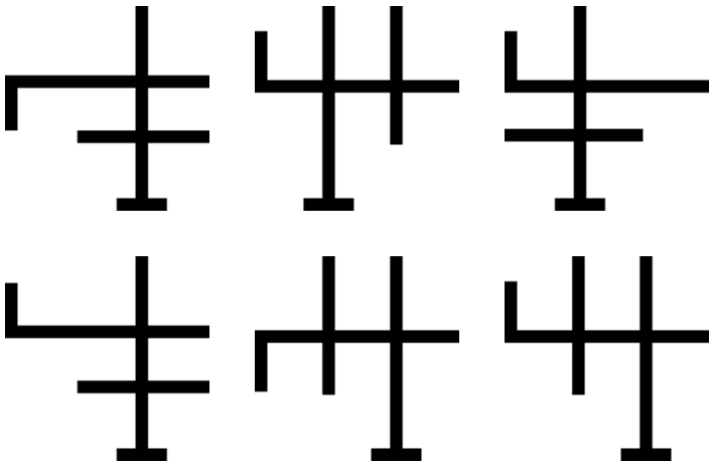

Figure 3

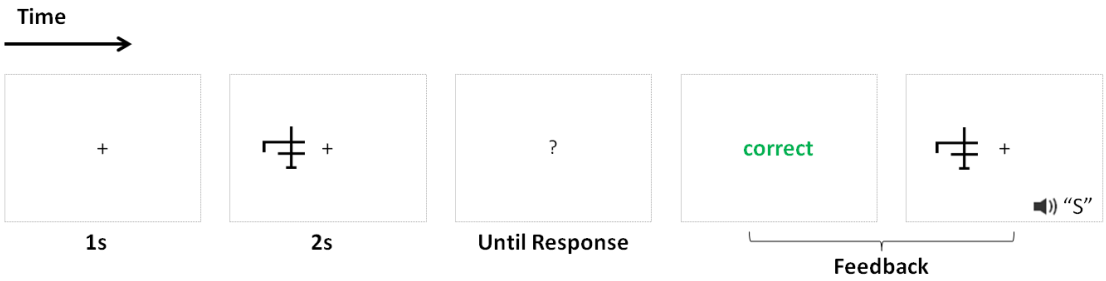

Figure 3. Schematic diagram of one training trail in Experiment 1. Participants fixated on fixation cross for 1 s and then a single object was displayed to for 2 seconds at a horizontal eccentricity of 5.5º in Experiment 1a and 6.5º in Experiment 1b. Participants attempted to retrieve the object name and then received feedback in terms of written response followed by the display of the object and its name presented auditorily.

Figure 4

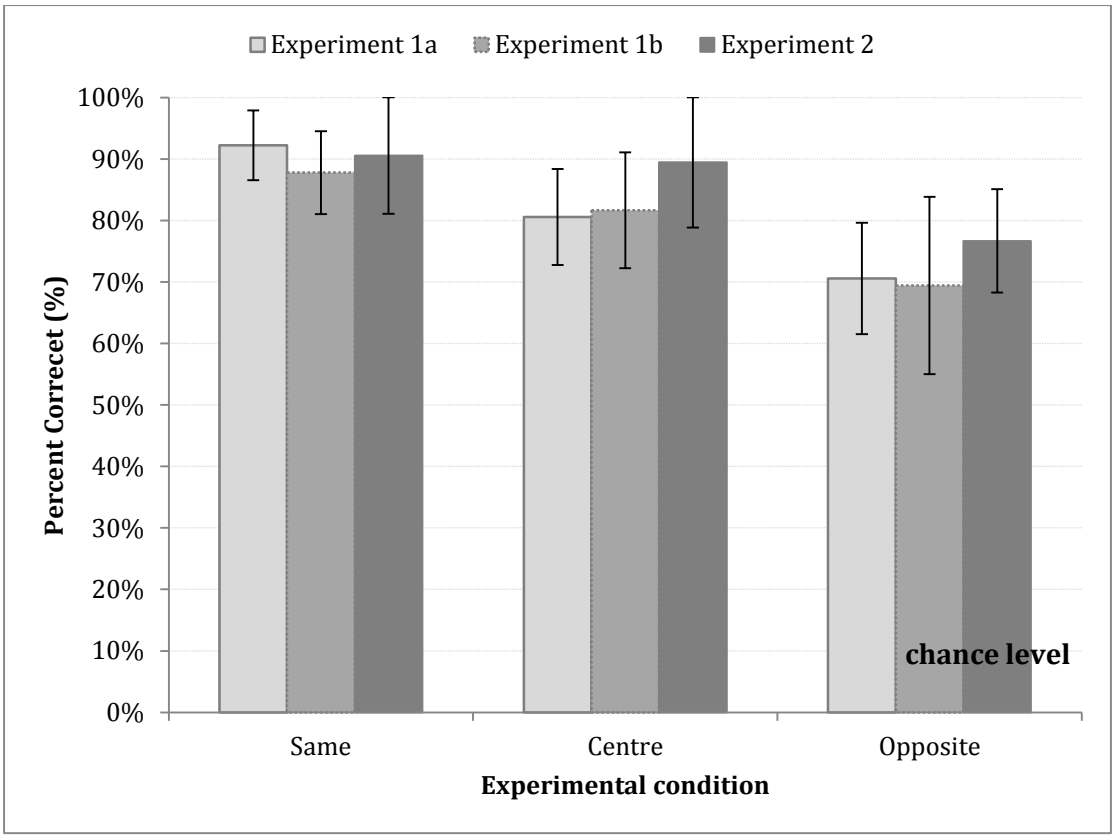

Figure 4. Percent correct object naming in Experiments 1 and 2 as a function of the study-test display conditions. Chance level is 16.7%. The error bars represent confidence intervals.

Figure 5

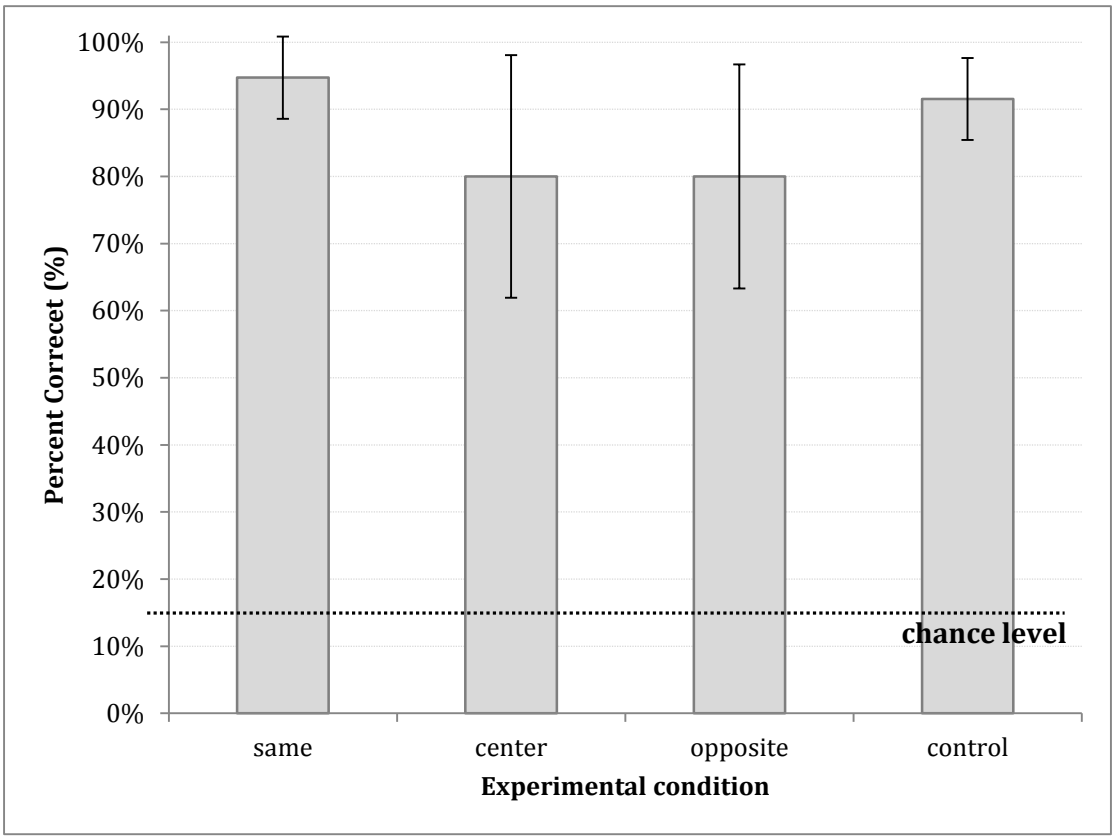

Figure 5. Percent correct object naming in Experiment 3 for the critical and control objects as a function of study-test conditions. Critical objects were studied in one location, control objects studied in all locations. Chance level is 16.7%. The error bars represent confidence intervals.